Document Classification with Reuter50 News Text: Proposal

Muzhe Zeng
mzeng6@wisc.edu

1 Dataset

The dataset is the subset of RCV1. These corpus has already been used in author identification experiments. In the top 50 authors (with respect to total size of articles) were selected. 50 authors of texts labeled with at least one subtopic of the class CCAT(corporate/industrial) were selected.That way, it is attempted to minimize the topic factor in distinguishing among the texts. The training corpus consists of 2,500 texts (50 per author) and the test corpus includes other 2,500 texts (50 per author) non-overlapping with the training texts. For more details, see this.

2 Algorithms

I plan to apply these three ML methods to perform document classification task.

- Logistic regression [1] with word counts and `vsp` [2] extracted latent factors.
- Support vector machine [3] with word counts and `vsp` extracted latent factors.
- Gradient Boosting Decision Trees [4] with word counts and `vsp` extracted latent factors.

I'll tune the number of latent factors, svm's kernel function + soft threshold, GBDT's tree depth + #features + iteration steps+ learning rate. I'll use cross-validation to choose hyperparameters and the test set for comparison between models. The evaluation will be the classification accuracy over the test dataset.

3 Timeline

November 17th Finish data processing work and start building at lease two ML models.

November 30th Build all three models and start fine tune the hyperparameters.

December 15th Wrap up experiment and writings.

December 17th Submission date.

4 Github Link

`https://github.com/MuzheZeng/Document-Classification-with-vsp`

References

[1] Hosmer Jr, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Applied logistic regression. Vol. 398. John Wiley & Sons, 2013.

[2] Rohe, Karl, and Muzhe Zeng. "Vintage Factor Analysis with Varimax Performs Statistical Inference." arXiv preprint arXiv:2004.05387 (2020).

2019 Spring High Performance Computing for Engineering Applications (ME 759) Final Project.

[3] Friedman, Jerome H. "Greedy function approximation: a gradient boosting machine." Annals of statistics (2001): 1189-1232.

[4] Cortes, Corinna, and Vladimir Vapnik. "Support vector machine." Machine learning 20, no. 3 (1995): 273-297.